RES MILITARIS
REVUE EUROPEENNE D ETUDES EUROPEAN JOURNAL OF MILITARY STUDIES
Social Science Journal

Artificial Intelligence based Computational Framework applied with DRL and CNN for Biometric Facial Authentication from Video Surveillance

Shivani Kumari, C S Raghuvanshi, Ankita Jain
Faculty of Engineering & Technology, Rama University, Kanpur, Uttar Pradesh, India
E-mail: shivani.kumari2k20@gmail.com, drcsraghuvanshi@gmail.com, ankitajain1691@gmail.com
*Corresponding Author: ankitajain1691@gmail.com

Abstract

Internet of Things (IoT) networks and authentication applications together are regarded as one of the most inventive technological models. Surveillance cameras and other IoT devices have captured live video footage in the outdoors. Use of facial biometrics in smartphones makes things secure and complex, which is known as system authentication. One of the difficult issues is accurately interpreting digital information in a natural setting. The noise effect introduces erratic fluctuations into the subject in the natural world, which could reduce the ability of the model to recognize objects. The proposed framework will identify facial biometric characteristics in an uncontrolled natural setting where non-symmetric random changes may exist. The goal of the proposed system is to reduce the false recognition rates (FRR) in unfavorable environmental circumstances. It transforms digital data with current surroundings of individual. The Viola-Jones algorithm has been used to determine the face region of interest (ROI) in the form of frames from a real-time surveillance dataset. The Deep Reinforcement Learning (DRL) algorithm has been developed for feature extraction from video frames. Goal of the proposed approach is to produce a binary tree that contains feature sequence. Further, the system establishes an association of feature sequences linked to the face identity using a convolutional neural network (CNN) model. The goal of the proposed work is to maintain good recognition of face biometric feature units in a realistic setting under different levels of randomization. Additionally, a variety of image processing assaults are evaluated on the proposed system. Finally, the achieved result is compared with five different state of the art techniques and the model achieves a 98.85% average accuracy rate.

Keywords: Video frame Processing, Viola-Jones Model, Deep reinforcement learning (DRL), facial unique biometric identification, convolution neural network (CNN).

1. INTRODUCTION

Even with present widely used and reliable biometric verification technology, a counterfeit might potentially create a duplicate certification by exploiting communication system vulnerabilities. To securely preserve their identities, users employ a type of intangible and unique identification known as biometric identity. An image is segmented into its constituent elements, or artifacts. For processing facial photographs, it is an essential tool, as Revathy et al. [1] explain. This study offers guidance on how to address the recognition imbalance brought on by irregular 1data creation. When face recognition is done dynamically in a high-dimensional workspace, where information sensitivity is paramount, there are numerous problems.

Face detection and identification seem to be useful in various video processing application domains, including user authentication, video forensic investigations, and human activity recognition [2]. Video data is often processed using face features from IoT devices, such as security cameras, for authentication tasks. Such an Internet of Things device can function in a natural environment and record digital data in real-time. The recognition rate may decrease if facial features are identified from video clips of poor quality. When feature qualities in models aren't sufficient to guarantee accurate identification, they quicken the attribute learning process. The proposed method makes use of CNN and deep reinforcement learning to tackle issues related to poor-quality real-time video data. In a surveillance video, face of a person can be observed from a variety of positions, views, and perspectives. The proposed approach leverages the concept of augmented reality to establish geometric relationships between the facial characteristics and the container environment. The several layers of CNN are capable of deciphering the low-quality video frames [3]. The CNN classifier correctly classifies facial photos using the collected attributes. Effective Face recognition from real time video clip is one of the issues that needs to be resolved [4]. As digital technology advances, security concerns about the veracity of digital media grow. For a variety of video processing applications, such as user identity authentication and digital forensic investigations, face recognition is helpful. It improves facial recognition in real time and in natural environments by utilizing the advantages of augmented reality. Augmented reality allows processing of digital information which may accommodated with real time changes to light, angles, attitudes, and other aspects of the physical world. If the model reads from a low-quality video clip, the identification rate can decrease [6]. When feature statistics cannot be used to determine the accuracy of recognition, the learning process may accelerate. The proposed model primarily makes use of CNN and deep reinforcement learning models to overcome the difficulties associated with identifying distinct facial features from footage captured in an unpredictable and turbulent environment. To enhance the quality of features, this system needs to perform prospective frame processing tasks, which are carried out by several CNN model layers. The convolutional model uses the collected variables to classify face photos and preserved recognition performance. Polarized video data might not produce consistent recognition accuracy [7]. This study maintains an excellent rate of detection in the face under numerous image processing assaults demonstrating the robustness of the system. A lot of study has been done on facial data classification. Deep reinforcement Learning (DRL) is method which is used for the feature extraction. This method involves learning of features from its feedback loop and tries to maximize the final reward in the form of binary tree which corresponds to facial unique traits. The CNN (convolutional neural network) model has been used to identify and classify the unique traits belongs of biometric traits of each individual capture in surveillance footage. The CNN model contains various hidden layers that effectively perform parallel processing of facial feature sequences in various segments and then it collaborates the output to the final layer. The degree to which recognition of images is used in picture breakdown into its component parts depends on the difficulty that has to be solved. Primary objective of decomposition is to divide an image into constituents that have a strong correlation with the regions of interest in the image [8]. There are two main kinds of categorization: partial and complete. A "filtering" strategy is employed when combined with the selection. If the criterion decides effectively, wrapper technique is employed in the attribute selection procedure.

The primary innovation of the proposed work is that, instead of using symmetrical static photos, it uses an Internet of Things surveillance video recorder to continuously recognize faces while employing video data in an unpredictable setting system. The framework makes use of real-time video data captured by cameras, blended with digital data and the natural surroundings. It shows augmented reality components and collects digital data while being impacted by external factors. The video in the proposed approach took inspiration from a widely used dataset [5] called chokepoint. This work successfully applies viola jones, deep reinforcement learning, and CNN principles to face recognition. The strategy aims to lessen any adverse effects that the surroundings might have on the ability to recognize face features. The proposed approach helps locate and recognize facial characteristics from input images in a variety of unexplored situations, including dim lighting, background noise, blurring, different positions, viewpoints, and altering expressions. Lower error rates (FAR and FRR) in the face of different computational image dangers are the goal of this work. This work validates the resilience of model to many tempering on the video frames while demanding rich input. Occasionally, multiple images of different people may appear in a single video frame, making it difficult for any basic system to identify each unique face independently. With the help of a natural environment, the simulation highlights the advantages of augmented reality and advances our understanding of the topic of real-time face trait identification. In a frame containing multiple faces, the suggested model performs exceptionally well and maintains a high level of efficiency. The arrangement of the additional modules is as follows: In Section II, the related work is illustrated. In Section III, the recommended methodology is presented. Section IV contains the results of the result presentation. The conclusion will be summarized in Section V. The final module contains a list of references.

2. RELATED WORK

In the past, convolutional neural networks have drawn interest for its ability to identify faces from image datasets by utilizing facial features. CNN appears to be competent and able to get a true correctness rate. A deep learning technique [9] has been employed to avoid noise-related issues such as distortion from the retrieved facial attributes of multiple people at once. This inflexible operational process of the study, which causes the framework to fail to maintain a uniform convergence, is one of its limitations. This method provided for the extraction of qualities to be combined separately and then the selection of the best features for exact mapping with each other in order to ensure the identification of unique facial identity. It is discovered that the efficiency as 92.3%. An alternative method [10] demonstrates how to identify facial features in asymmetric environments, like ones with shifting lighting, blurring, and poor resolution. The reliability of the model with symmetrical datasets also its limitation, as any changes to the dataset may cause the model to lose its high level of robustness. The Gabor filter was utilized by CNN to conduct face recognition while handling data that could have unpredictable effects. The method was using a vector of characteristics to map Gabor waves into the Euclidean space. A other approach [11] combines principal component analysis, local binary patterns, and a multi-feature extractions algorithm. These two feature extraction techniques yield the best characteristics that together make up CNN input information. The accuracy of this procedure is found to be 90%. The inherent temperedness of the dataset is the reason for this limitation in terms of generalization and interpretation of hidden feature sequences. A different study [12] extracted the 1D features of the face constituents from the image collection using a linear discriminant analysis technique. A DNN classifier with 92% accuracy can be used with the attributes. It has been found, therefore, that the algorithm is less trustworthy and may not be able to recognize the correct class of photos with atypical noise. A method [13] provided a model that handled illumination normalization as the pre-processing step and conducted segmentation of the image based on neighboring pixels using a reformed FCM method. This division efficacy is higher and more noise-tolerant than other FCM-focused techniques. The prediction of classifier has been determined in relation to the selection of the distinctive feature assessment criteria [14].

3. PROPOSED METHODOLOGY

The proposed operation of model for a face recognition job using video data is depicted in Figure 1. The proposed method gathers footage from IoT monitoring system in an ambient environment by placing a camera inside a room and recording different persons approaching to it.

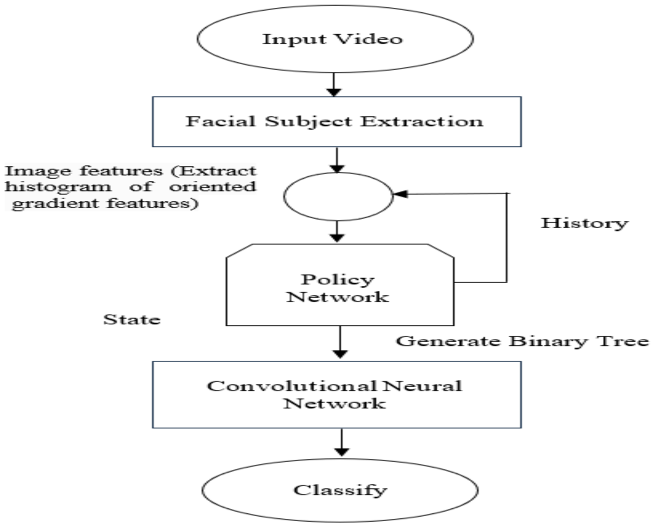


Fig. 1. Working of Face Recognition System

Video consist of many frames, each frame of the video captures data contains several images of the same person. These clips are then analyzed in order to identify the faces of users. Figure 2 illustrates some video frames [5].

Fig. 2. Sample images from Video Frames [5]

As seen in the Figure 2, a great number of people have been captured in video frames. The image size is 552 by 480 pixels, and the frame rate is set to 30 frames per second.

3.1 PRE-PROCESSING

An efficient real-time object recognition method might be able to extract a portion of the face of subject from the incoming footage. Viola Jones [14] has been applied in the proposed solution. The method, which looks for distinctive features in facial data, is employed by the model to identify certain regions in different video frames. These characteristics come combined and remain unchanged in the recorded video frame despite of background noise, poor lighting, strange facial expressions, and posture of an individual. The technique looks at every feature point in the video frames to identify the face of a subject. Further inscribes it in gaps with a rectangular form. A sample of a frame produced by the Viola-Jones algorithm is shown in Figure 3. The individuals depicted in Figure 3 are those who willingly took part in the experiment to create the chokepoint dataset, which is accessible to the public [5].

Fig 3. Extracted Faces using Viola-Jones algorithm

$$\text{Image(P)}= \sum_{i=0.0}^{i} \sum_{j=0.0}^{j} yz$$

(1)

P, the group of favorable windows that cascade

$$W = \{[i,i+a-1]\} * [u, u+a-1] € L$$

(2)

$$W_i = \frac{w_i}{\sum_i w_i}$$

(3)

Equation 1 shows the input frame which is the part of 2 matrices of y and z. The multiplication of comparable feature vectors from input video frames is displayed in Equation 2. Drawing the haar traits across the frame illustrates how the concentric boundary surrounding the facial segment is processed. The normalization of the segmentation facial subject in the outer region is presented in Equation 3.

3.2 DEEP REINFORCEMENT LEARNING

A deep reinforcement learning system is used to learn from its individual input. In DRL, HOG features are extracted from the collected facial information using the Viola-Jones algorithm and then converted into a binary tree-shaped architecture. Recurring features from many face photographs of the same person are utilized to generate binary codes that are comparable to each other. Every face segmentation feature was mapped to a similar binary tree. The proposed approach will record a single individual arriving repeatedly in frames of video. In this instance, DRL is used to build a new binary tree for every frame, which usually contains the same person who appears several times. The technique is designed to create the same number of trees for every human being that is filmed. As a result, each binary layout only designates one linked topic that is distinguishable from the other facial biometric binary layouts. The proposed method uses an agent that performs hashing to create fixed-size binary codes from hog characteristics. The hash codes have been developed in order to maximize the reward. Binary codes are prone to mistakes. The DRL model has employed a back-propagation strategy to transmit the errors back to an earlier state where the weights of the agent function has been modified to produce the optimal reward. Binary codes corresponding to the closest similar structures have been associated with single-person image sequences. Ultimately, the CNN algorithm utilizes binary traits in order correctly identify the photos and recognize the biometric identity uniquely.

A policy network, at the heart of DRL system, builds a binary tree through trial-and-error exchanges. The state, which includes the primary characteristics of the face image, is used to translate the original picture of the face to comparable hash codes. Since there may be initial defects in the state, a feed-forward network often instructs the agent employing an earlier selection that may have errors. The network computes the descent of a loss function in order to keep the weights in balance. The grade of the generated binary codes may be confirmed using the reward function. Maximizing the total reward value is the main goal of deep reinforcement learning [15]. Equation 4 states that the $\omega(A_t)$ symbol represents the normal exponential function of each input vector element. The input feature number is denoted by "m". Both computing time and memory use are reduced.

$$\omega(\overrightarrow{A_t}) = \frac{e^{A_t}}{\sum_{m=1}^{n} e^{A_t}} \quad for\ t = 1,2,3,\ldots\ldots.n \tag{4}$$

$$Loss = -\sum_{p=1}^{q} z_i \log(\widehat{z_i}) \tag{5}$$

$$Loss = -\log(\widehat{z_i}) \tag{6}$$

The following are the elements of the loss function used in Equations (5) and (6): z_i represents the probability that a category actually exists, p stands for the overall number of classes, and $\widehat{z_i}$ represents the chance that a class is expected to exist. The DRL agent determines the likelihood of action for a given state. The only digits in the hash or binary codes that are produced are 1 and 0. The probability distribution used by the agent to forecast the hash codes for the current state.

3.3 CONVOLUTIONAL NEURAL NETWORK MODEL

The system uses a neural network to apply biases and weights to the input of a signal while putting individual neuron connections in order. Hidden layers have been connected using the same technique. The model governs functionality by changing the weights and quantifying many functions [16]. CNN model gains an advantage for handling integrated scenarios, especially in limited design architecture. If the system is unable to identify a pattern after the data has been supplied from the input, it automatically responds by employing many levels that are equivalent to certain values multiplied by 0 or 1. Figure 4 shows how selecting an image from a video helps with feature extraction. Feature comparison is applied to each and every pixel in the picture [17]. Then apply point function and its primary duty is to establish a relationship between the details and the data.

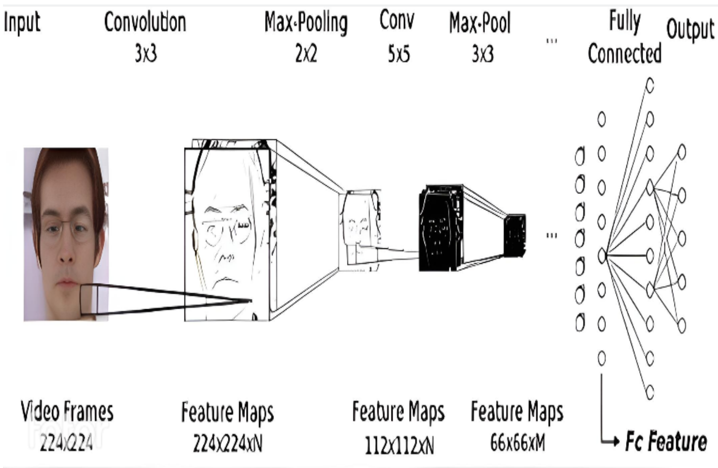


Fig. 4. Working of the CNN model

Figure 4 shows the number of processing modules employed in CNN. It contains convolutional layer, Max-pooling layer, again convolutional layer of different size, max-pool layer of variant size, and at the last fully connected layer has been employed to generate the final output.

4. EXPERIMENTAL SETUP AND RESULTS

4. 1 EXPERIMENTAL SETUP

To conduct the experiment, MATLAB tool has been used of version 2021b. The setup includes CNN supporting MATLAB libraries. The GUI simulation of the working model has been constructed on MATLAB software. The setup contains 8 GB of RAM, 2 GB NVIDIA graphic card and 500 GB of SSD drive to support the hardware infrastructure of the working model.

4.2 DATASET

The dataset used in the proposed work is called the chokepoint dataset [5]. It has been collected by ICT Center of Excellence program (Australia). In order to train the model, a total of 49 videos featuring the sequences of 62,324 pictures from VGG FACE [16] have been captured using IoT cameras. It contains the 244 × 244 input image size and the 228 classes. The dimensions of the completely connected layer are 4098. The static framework is located in the leftmost section of image. Each of these two state dimensions is 224 × 224 x 3. The system generates five three-dimensional frames: three 112×112×128 frames and two 224×224×64 frames. Similarly, three-dimensional frames are chosen for the next phase.

4.3 RESULTS

Multiple people entering in the room in 126 recorded frames that are collected by the video recorder. For the camera, each person is photographed in five different positions.

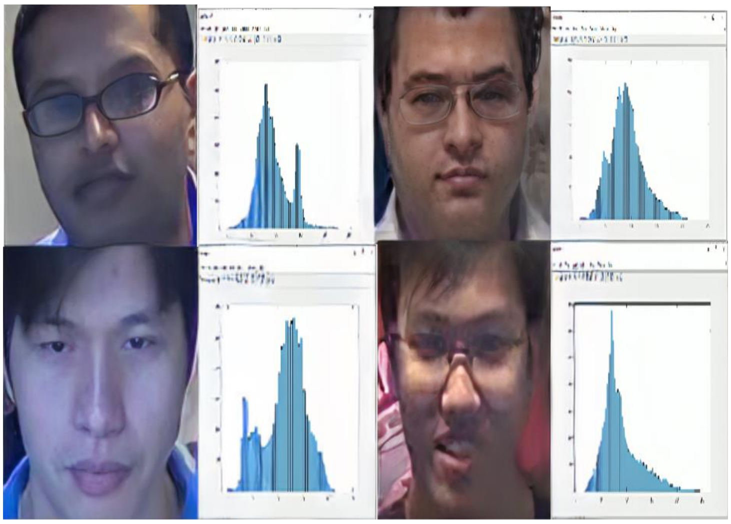
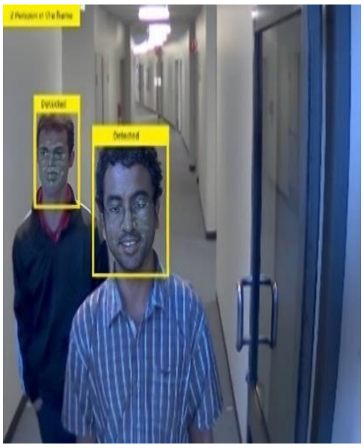


Fig. 5. Histogram representation of Number of people detected by the model

Figure 5 shows how facial features are represented as a histogram. The Viola Jones algorithm receives these pre-processed images and uses them to extract facial features. Subsequently, features have been extracted using reinforcement learning. Figure 6 displays the extracted facial subject, a plot of histogram feature points, HOG feature visualization, and the relationship between acquired video frames with the same person type. The model clearly shows that each person was identified uniquely with the use of a statistical representation. Consequently, every individual may be identified by the system from video clips containing crowd capture. These features of the spectrum are employed in training.



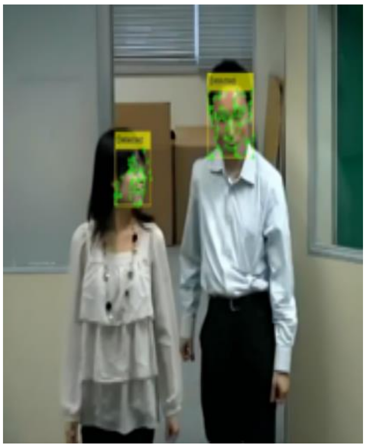
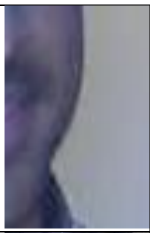
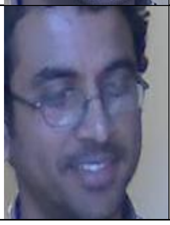


Fig. 6. Image samples with Identified region of interest

Figure 6 displays the facial components that were recognized using the proposed technique in the security camera. A rectangular block encircling the face content has been plotted by the viola Jones method using a feature plot (green dots - HOG features). To assess the robustness of the model, several image processing techniques are performed to video frames; the results are shown in Table 1. The associated peak signal-to-noise ratio (PSNR) and normalized cross-correlation (NC) are calculated in order to evaluate the video frames. The effects of image tempering are limited to a single face in Table 1.

Table 1. Implication of tempering on video frames

Types of Attacks	**Original Images**	**Tempered Images**	**Normalized Cross Correlation**	**PSNR (dB)**
Noise			0.9412	92
Rotation 45 degree			0.9342	95
Shear			0.9672	96

Crop			0.9231	94
Deviation in Mean			0.9543	97

After the impact of multiple tempering, the average PSNR value is 96 dB and mean Normalized correlation is 0.9146. These findings show that the system is resistant to several applications of tempering with frames. The study and effective results show the strength the proposed paradigm. Consequently, the model may still be able to precisely and flawlessly recognize the facial features even in the situation of tempering. In Figure 7, the ROC curve is displayed, providing an overall picture of the accuracy of the proposed system.

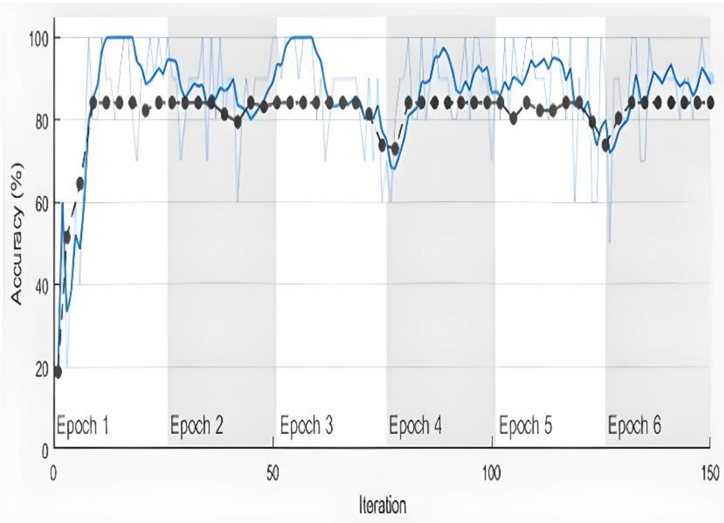


Fig. 7. Accuracy by ROC curve

The proposed mean precision was found to be 98.85%. A number of photo examples are utilized to evaluate the model. The bottom part of Figure 7 also has the lowest error rate. Results of figure 7 indicates that the model performs better when real-time photos are taken in a variety of settings, such as with changing lighting, expression, posture, etc.
The CNN functions are not linear. It states that the multiplication of the input characteristics and weights with the addition of a bias input value is what initiates the operation of perceptron.

Fig 8. Performance assessment using Confusion Matrix

Figure 8 shows the results of the efficiency evaluation using the confusion matrix, which was computed during the gathering of video data with six distinct person counts. It is clear from the above graphic that the system can continue to operate at high levels even in the presence of unprecedented noise effects. Table 2 compares the proposed system with existing technology.

Table 2. Comparison of proposed work with state-of-the-art techniques

S. No.	Techniques	Accuracy (%)
1	Basil et al. [12].	91.60
2	Alhanaee et al. [13].	84
3	Singh et al. [14].	96.44
4	Srivastava et al. [15].	78
5	Huber et al. [16].	95.71
6	**Proposed Work**	**98.85 (Attained)**

5. CONCLUSION

The proposed method, which uses video feed from an IoT security camera positioned in a natural environment, is a dependable and efficient way to identify faces. To achieve this goal, the system makes use of the CNN architecture, the DRL method, and the Viola-Jones algorithm. For real-time image samples, the highest identification accuracy of up to 98.85% is attained. Multiple image tempering are also applied to the different returned images to test the ability of model to identify the corresponding photos. It looks at the challenges of creating augmented reality in a realistic environment with an unprecedented variety of lighting instances on frames. It is also found that the results are dependable and acceptable. The model is able to recognize all of the faces when they are displayed in a single video frame. The strength of the proposed model is in its ability to maintain high accuracy even in the face of the extraordinary variation brought about by the natural world. The model operates on a dynamic, random dataset that is subject to random noise challenges in real time. Future research will examine the tempered photos in further detail in a variety of ways.

REFERENCES

[1] Revathy, G., Raj, K. B., Kumar, A., Adibatti, S., Dahiya, P., & Latha, T. M. (2022). Investigation of E-voting system using face recognition using convolutional neural network (CNN). *Theoretical Computer Science*, *925*, 61-67.

[2] Ikromovich, H. O., & Mamatkulovich, B. B. (2023). FACIAL RECOGNITION USING TRANSFER LEARNING IN THE DEEP CNN. Open Access Repository, 4(3), 502-507.

[3] Lazarini, M. A., Rossi, R., & Hirama, K. (2022). A systematic literature review on the accuracy of face recognition algorithms. EAI Endorsed Transactions on Internet of Things, 8(30), e5-e5.

[4] Hung, B. T. (2021). Face recognition using hybrid HOG-CNN approach. In Research in Intelligent and Computing in Engineering: Select Proceedings of RICE 2020 (pp. 715-723). Springer Singapore.

[5] The chokepoint dataset is sponsored by NICTA. NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy, as well as the Australian Research Council through the ICT Centre of Excellence program, http://arma.sourceforge.net/chokepoint/.

[6] Gwyn, T., Roy, K., & Atay, M. (2021). Face recognition using popular deep net architectures: A brief comparative study. Future Internet, 13(7), 164.

[7] Lu, P., Song, B., & Xu, L. (2021). Human face recognition based on convolutional neural network and augmented dataset. Systems Science & Control Engineering, 9(sup2), 29-37.

[8] Tabassum, F., Islam, M. I., Khan, R. T., & Amin, M. R. (2022). Human face recognition with combination of DWT and machine learning. Journal of King Saud University-Computer and Information Sciences, 34(3), 546-556.

[9] Boussaad, L., & Boucetta, A. (2022). Deep-learning based descriptors in application to aging problem in face recognition. Journal of King Saud University-Computer and Information Sciences, 34(6), 2975-2981.

[10] Gupta, S., Thakur, K., & Kumar, M. (2021). 2D-human face recognition using SIFT and SURF descriptors of face's feature regions. The Visual Computer, 37, 447-456.

[11] W. Deng, J. Hu, and J. Guo. Face recognition via collaborative representation: Its discriminant nature and superposed representation. IEEE Trans. Pattern Anal. Mach. Intell., PP(99):1–1, 2018.

[12] Basil, N., Raad, M., Wazzan, A. N., & Marhoon, H. M. (2022). Face recognition with real-time framing based on multi task convolutional neural network: a case study. Int. J. Mech. Eng, 7(2), 3170-3178.

[13] Alhanaee, Khawla, Mitha Alhammadi, Nahla Almenhali, and Maad Shatnawi. "Face recognition smart attendance system using deep transfer learning." Procedia Computer Science 192 (2021): 4093-4102.

[14] Singh, A., Prakash, S., Kumar, A., & Kumar, D. (2022). A proficient approach for face detection and recognition using machine learning and high-performance computing. Concurrency and Computation: Practice and Experience, 34(3), e6582.

[15] Srivastava, S., Kumar, A., Singh, A. et al. An improved approach towards biometric face recognition using artificial neural network. Multimed Tools Appl 81, 8471–8497 (2022). https://doi.org/10.1007/s11042-021-11721-2.

[16] Huber, M., Luu, A. T., Boutros, F., Kuijper, A., & Damer, N. (2024). Bias and Diversity in Synthetic-based Face Recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 6215-6226).

[17] Yang, X., Liu, C., Xu, L., Wang, Y., Dong, Y., Chen, N., ... & Zhu, J. (2023). Towards effective adversarial textured 3d meshes on physical face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 4119-4128).

[18] H. Kumar, P. Mamoria, D. K. Dewangan and S. Kumari, "Internet of Things (IoT): Role of Machine Learning in Power Management and Optimization using Regression Analysis," *2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT)*, Greater Noida, India, 2024, pp. 1319-1326, doi: 10.1109/IC2PCT60090.2024.10486680.

[19] H. Kumar and A. Singh, "Internet of Things: A Comprehensive Analysis and Security Implementation through Elliptic Curve Cryptography", *International Journal of Current Engineering and Technology*, vol. 6, no. 2, pp. 498-502, 2016.

[20] Kumar, H., Mamoria, P., Kumari, S., Misra, M.K. (2024). Video Steganography Techniques: A Comprehensive Review and Performance Evaluation. In: Chaturvedi, A., Hasan, S.U., Roy, B.K., Tsaban, B. (eds) Cryptology and Network Security with Machine Learning. ICCNSML 2023. Lecture Notes in Networks and Systems, vol 918. Springer, Singapore. https://doi.org/10.1007/978-981-97-0641-9_3